

A METHOD FOR DETECTING NEW CONTENT IN A TARGET WEB PAGE WITH REFERENCE TO A CONTROL WEB PAGE USING NORMALIZED TEMPLATE MATCHING

ABSTRACT

A method called normalized template matching is described which detects new content in any given target web page with reference to a control web page. The control page is typically an older version of the given target web page. The control page is first divided into sections by using non-formatting HTML tags. Each section is then individually normalized by removing formatting HTML tags, meta characters and repetitive white spaces, after which they are inserted into a template in sorted order. Each entry in the template consists of the section content and its type. Once the entire control page is processed, a normalized template is obtained. The target page is similarly divided and normalized, after which each section of the target template is matched against the entries in the control template. Any section of the target page which does not match the control template will be flagged as new content and presented in a summary page to the user.

